Recruit Until It Fails: Exploring Performance Limits for Identification Systems

SHRIDATT SUGRIM, CAN LIU, and JANNE LINDQVIST, Rutgers University

Distinguishing identities is useful for several applications such as automated grocery or personalized recommendations. Unfortunately, several recent proposals for identification systems are evaluated using poor recruitment practices. We discovered that 23 out of 30 surveyed systems used datasets with 20 participants or less. Those studies achieved an average classification accuracy of 93%. We show that the classifier performance is misleading when the participant count is small. This is because the finite precision of measurements creates upper limits on the number of users that can be distinguished.

To demonstrate why classifier performance is misleading, we used publicly available datasets. The data was collected from human subjects. We created five systems with at least 20 participants each. In three cases we achieved accuracies greater than 90% by merely applying readily available machine learning software packages, often with default parameters. For datasets where we had sufficient participants, we evaluated how the performance degrades as the number of participants increases. One of the systems built suffered a drop in accuracy that was over 35% as the participant count increased from 20 to 250. We argue that data from small participant count datasets do not adequately explore variations. Systems trained on such limited data are likely to incorrectly identify users when the user base increases beyond what was tested. We conclude by explaining generalizable reasons for this issue and provide insights on how to conduct more robust system analysis and design.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Supervised learning by classification**.

Additional Key Words and Phrases: machine learning, identification, sample size, quantitative methods

ACM Reference Format:
Shridatt Sugrim, Can Liu, and Janne Lindqvist. 2019. Recruit Until It Fails: Exploring Performance Limits for Identification Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 104 (September 2019), 26 pages. https://doi.org/10.1145/3351262

1 INTRODUCTION

Identifying individuals is a key component in many systems like automated grocery (e.g. Amazon Go [2], Alibaba Taocafe, DeepBlue Takego), personalized recommendation systems [41] (e.g. ads [16, 54], movies [36], products [3], music [51]), or multi-user interfaces [14]. The typical identification system measures observable features of a user and then feeds these measurements to a decision mechanism. The decision mechanism learns the distributions of measurements from the dataset and makes predictions by partitioning the space of measurement values.

This work is supported by the National Science Foundation under Grant Number 1750987. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional material available at http://scienceofsecurity.science.
Authors' address: Shridatt Sugrim, shridatt.sugrim@rutgers.edu; Can Liu, can.liu@rutgers.edu; Janne Lindqvist, janne.lindqvist@rutgers.edu, Rutgers University, 94 Brett Road, Piscataway, New Jersey, 08854.

2474-9567/2019/9-ART104 $15.00
https://doi.org/10.1145/3351262

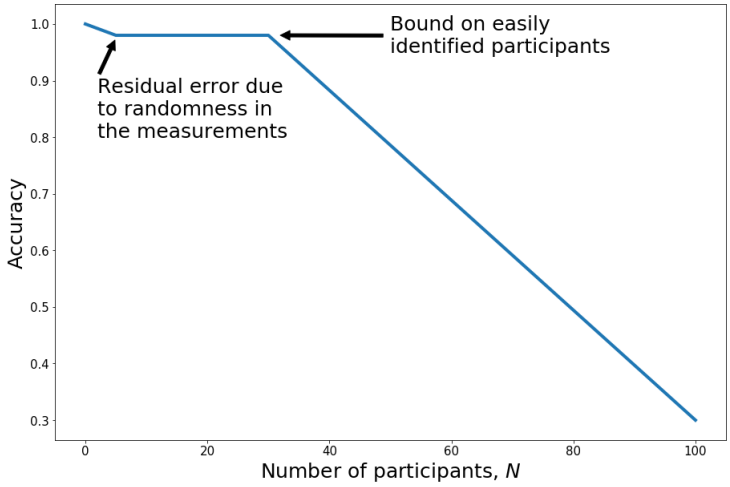

Fig. 1. Even an ideal system that is capable of identifying many users with minimal error will eventually reach an upper bound, $\mathcal{N}$, beyond which the performance will decrease with out recovery. Because the measurement value range is finite, the ability to distinguish participants based on these measurements begins to degrade as more participants are added. Consequently, users that were easily identified will be confused with new users that have similar measurement values. We can choose a threshold (e.g. accuracy greater than 80%) below which we declare that the systems error rate makes it unsuitable. Beyond this point, we declare that the system has failed. Because the measurements can vary randomly, it is often very difficult to identify the bound, $\mathcal{N}$, beyond which the performance decreases monotonically.

Many recently proposed identification systems use machine learning classification algorithms as their decision mechanism [5, 13, 22, 24, 26, 28, 31, 33, 34, 38, 45–47, 49, 52, 57, 59–67, 69]. An identification system's performance is measured by how well it performs classification. These systems are often validated with a user study where participants are recruited, observables are measured multiple times, and then the measurements are used as a dataset for a classification algorithm.

For a system to be robust, it is critical to know what conditions cause it to fail. The user studies used to evaluate these proposed systems should provide some insights about the limits of the proposed systems. Even the most ideal systems which are capable of identifying large numbers of participants with minimal error will fail in unexpected ways when used beyond its upper bounds (an upper limit on number that can be identified with minimal error) as described in Figure 1. Often these bounds are not included in the analysis of a proposed system nor are they reported in the publications if known.

User studies reported in the literature are often inadequate to consider a system well tested. Although researchers may collect large amounts of measurement data per participant, the participant count is often low. Collecting a large number of measurements from a small group of participants does not test the limits of the decision mechanism. To verify that small participant counts is an issue for identification systems, we surveyed 30 recently proposed identification systems and noted that the median number of participants for their user studies was 12. Of the systems surveyed no system reported a limit on the number of users that can be handled by the system or identified the conditions which drive the system to failure. Several of these systems use classification algorithms such as support vector machines, neural networks or random forests as multiclass classifiers. Their performance is often measured using two common metrics, accuracy and the confusion matrix [53]. Accuracy is the relative frequency of a correct classification. The confusion matrix is a contingency table that enumerates

how often any one class gets confused for any other class. Together, both metrics quantify different aspects of how often the decision mechanism fails.

In this paper, we aim to make a strong generalizable claim: **no one should be surprised that classification algorithms are able to distinguish classes derived from small numbers of participants.** We demonstrate this by building five different identification systems based on five different publicly available datasets that measured humans. We tasked these systems with identifying the humans who were measured. We chose human-generated data to ensure the probability distributions of the data would be similar those encountered when conducting a user study to evaluate a proposed system. The datasets were used only as a source of human-generated measurements: we are not concerned with the datasets' original purposes. We purposely minimized the effort to make the classifiers perform better than guessing. We examine the reasons why this is easy to achieve. We consider the impact of *measurement count* (the total number of measurements taken across all participants), *feature dimension* (a function of number of distinct measured observables), *sample diversity* (how distinct the measurements of each participant are from each other), and *participant count* (number of participants in the user study). We use the insights from our dataset analysis to reason about the degradation of performance as the participant count increases. The major contributions are as follows:

We examined the recognition performance of identification systems with low participant counts. We mimicked the evaluation procedure of an identification system by constructing five distinct systems based on publicly available datasets, each with 20 participants. We used support vector machines, random forests, and neural networks on these datasets and achieved greater than 90% accuracy in three cases.

We explored the reasons why the low participant count is not adequate to evaluate identification systems. We examined the impact of participant count on performance results and observed that performance may degrade as participant count increases. In addition to the participant count, we also analyzed the impact of measurement count, number of measurements per participant, and feature dimension on performance.

We outlined a method for more rigorous testing of an identification system. We proposed measuring how the performance metrics degrade when the participant count increases as a gauge of the robustness of an identification system. We examined how to anticipate this degradation by using randomized participant subsets, and noted it as a crucial criteria to demonstrate the performance of a novel identification system.

It is fairly well accepted that having a small participant count in a user study is inadequate to asses that system [29, 44]. However, the literature does not document the impact of small participant counts on the performance metrics of identification systems proposed in literature. To determine the scope of this problem, we describe our survey of the literature on recently proposed systems in Section 2. We then construct five systems from human generated datasets in Section 3, which explore what performance can be achieved with minimal data processing effort. We then examine how the properties of a dataset impact the performance metrics in Section 4. From this analysis, we identify probability distributions that play a central role in system performance. We discussed how these probability distributions create bounds on the number of easily identified participants in Section 5. We argue that the bounds are a realization of limits on measurement precision. Because of the stochastic nature of these systems, it is hard to definitively identify when the bound has been reached. We suggest strategies for testing the robustness of a system without concrete knowledge of its bounds in Section 6. Finally, we give related work in Section 7 and summarize the findings in Section 8.

Table 1. Publications surveyed grouped by publication venues. Others category includes Infocom, MobileHCI, and MobiSys

Venue	Case 1: User Identification	Case 2: Multiclass Identification
IMWUT/UBICOMP	[55], [49], [67], [26], [61], [62], [13]	[31], [52], [63], [69], [57], [65], [43]
CHI	[46], [38], [45]	[66], [39], [64], [47], [34]
UIST		[5], [22], [33]
OTHERS	[24]	[28], [56], [60], [59]

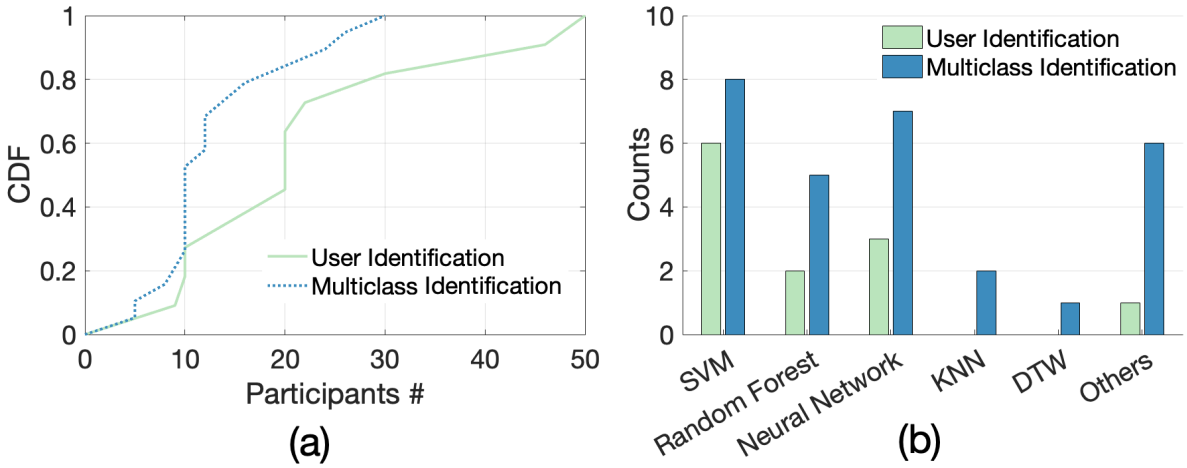

Fig. 2. Summary of the participant counts and classification approaches used in the surveyed two types of identification systems: user identification and multiclass identification. (a) shows the cumulative distribution function of the two types of identification systems. We found more than 77% of the publications in these two types of systems recruited 20 or less participants in their user studies. (b) summarizes the classification techniques used. We found support vector machines (SVM), random forests and neural networks are among the most common. Others category includes Hidden Markov Model [39, 55], Jaccard similarity coefficient [56], k-nearest neighbors [52, 65].

2 SURVEY OF RECENTLY PROPOSED IDENTIFICATION SYSTEMS

To understand how many participants were recruited in recent publications, we surveyed the papers published in top venues during the last four years (2016-2019). We focused on systems literature where machine learning is often used as a black box. We considered systems papers from top-tier conferences in mobile and ubiquitous computing, human-computer interaction and networking. These conferences included CHI, IMWUT/UBICOMP, Infocom, MobileHCI, MobiSys, MobiCom (no papers discovered), and UIST (see Table 1).

The systems proposed in these articles can be separated in to two cases: multiclass identification systems and user identification systems. A multiclass identification system measures a user and attempts to predict one of several classes (e.g. standing posture, handwritten digits, or a hand gesture in free space). The user identification system is a subset of the multiclass identification system where the classes are in one-to-one correspondence with the users. That is, each class uniquely identifies an individual user. For multiclass identification, the number of classes could be less than the number of participants (e.g. a fixed set of gestures). In this case the probability distributions of measurements become concentrated into a smaller set of classes. The underlying decision mechanism is still the same in both cases because the distribution of classes depends on the distribution of measurements. These systems measure users, compute distributions of classes based on those measurements,

Table 2. We chose datasets that had at least 20 participants. The total measurement count is often misinterpreted as the sample size. The average number of measurements per participant can indicate how well-characterized the statistics of an individual participant is. The feature dimension is an indication of the complexity of what is being measured.

	Act. Recogn.	Walking Act.	CAT Scan	EEG	NBA Stat.
Participant Count	30	22	97	81	296
Total Measurements	10299	149332	53500	23986	5444
Average Measurements	343	6788	552	296	18
Feature Dimensions	562	4	385	39	51

and then predict on new inputs based on those distributions. We argue that the reliability and performance of identification systems cannot be fully evaluated with the user studies of low participant counts.

We did not consider machine learning and data mining conferences, such as NeurIPS and KDD, because they focus on algorithms instead of system applications and often use curated datasets instead of generating their own data by recruiting participants. Figure 2 summarizes the participant counts and techniques that were used in the 30 surveyed publications. Our goal is to bring attention to the misleading results that arise from recruitment practices and to advocate for testing for failures.

3 DATASETS AND CONSTRUCTION OF THE IDENTIFICATION SYSTEMS

To analyze the potential issues of the identification systems with small participant counts, we constructed five user identification systems using publicly available datasets. We chose the user identification task as it was easy to implement, had readily available data, and has a simple interpretation of the measurement distributions. We used common classification techniques with minimal tuning. We discovered that high classification performance is achievable when the participant count is low.

In all cases, the data was used only as a source of human-generated measurements. We did not assess the datasets usefulness for its collected purposes (e.g recognizing walking activity). We were solely interested in the discerability of the measurements from distinct participants. Our study was approved by the Institutional Review Board (IRB) of our institution.

3.1 Choosing Datasets

To identify datasets that we could use, we examined datasets from several public repositories including, UCI Machine Learning Repository [15], Kaggle.com [21], Data.gov [35], and other public data repositories. For a dataset to be included in our study, the dataset had to meet the following simple criteria:

- Unique identifier for each participant
- More than 20 participants
- More than one measurement per participant

We did not restrict datasets based on measurement type, number of features, or other dataset properties to maintain generalizability of the results. Table 2 lists the datasets.

3.2 Choosing Identification Methods

To construct a user identification system, we applied the three most common algorithms used in our survey reported in Section 2 - Figure 2(b)): random forest, support vector machines and neural networks. Each identification system was constructed with 20 users with unique identifiers. The unique identifiers served as the class labels that would be learned by the machine learning algorithms. The systems were evaluated on how well

they predicted the identifier when given an unlabeled measurement. In the case of a multiclass identification system the evaluation would be the same, however, the number of classes may not be equal to the number of participants.

All classifiers were implemented using the sci-kit learn [37] library. We aimed to minimize the amount of machine learning knowledge required to implement an identification system with high performance. We performed very little optimization on each of the algorithms. In the cases where we did not use the default parameters, the selected parameters were chosen strictly to prevent infinite loops and minimize run time in order to treat the machine learning as a black box. Our principal effort was to build identification systems from human-generated data. Our goal was to demonstrate how misleading results will arise when a selection of participants produced a dataset that was artificially easy to classify.

3.3 Identification Systems Performance

The performance metrics we computed were 1) the accuracy score (ACC), 2) the confusion matrix (CM), and 3) the number of easily identified users. The first two metrics are widely reported in our survey of identification systems and machine learning literature when multi-class classification is evaluated [5, 12, 13, 22, 24, 28, 31, 33, 34, 39, 45, 47, 56–62, 64–66, 68]. The number of easily identified users is a simple metric derived from the confusion matrix (see table 4).

These metrics were computed using a standard validation technique where the data is split into two subsets, a training set which consists of 80% of the measurements and a testing set with the remaining 20%. For datasets where there were more than 20 participants, we also ran the analysis over multiple randomly chosen subsets of 20 participants. This was done to eliminate the possibility that a specific chosen subset would inflate the performance metrics purely by chance. A large variation in performance across these randomized subsets would indicate bias in the identification system.

No preference was given to any particular algorithm. In the case of support vector machines and neural networks, the default parameters were used. For random forest we set the n_estimators (number of decision trees to test) to 1000 to ensure a large breadth search and the max_depth to 20 to prevent infinite loops (more details in Appendix B).

3.3.1 Accuracy Score. The accuracy score provides a simple summary of performance by computing the relative frequency of a correct decision. This summary, however, is incomplete as it looses details of the systems performance on each individual participant. Additionally, the accuracy score can be misleading when the number of measurements per participant is unbalanced [53].

To calculate the accuracy score (ACC), we ran ten iterations with randomized participant subsets and reported the best accuracy achieved across all sets. Table 3 shows the best results achieved across ten different resamplings. *This method of model selection demonstrates a scenario where the performance is misleading because of the serendipitous favorability of the dataset.* In one case the reported accuracy was 100%.

The classifiers performances were similar when the measurement values from different participants was very distinct. In all cases, the accuracy metric varied by at most $\approx 8\%$. The values presented in Figure 3 are among the highest observed to highlight what is achievable with favorable subsets. Table 3 shows that three of the data sets had at least one algorithm that was able to achieve 80% accuracy. *Thus, an algorithm was discovered that would achieve reasonable performance with minimal tuning.*

All algorithms performed poorly on the EEG readings dataset regardless of parametrization of the algorithms. The best achieved accuracies were $\approx 52\%$ for random forest, $\approx 51\%$ for neural networks and $\approx 33\%$ for support vector machines. Parameter optimization techniques such as grid / random parameter search did not improve the results. This observation informs the strategy we propose and we will discuss why the reasons for this failure lie with the measurements in Section 4.

Table 3. Maximum Accuracy - We achieved greater than 50% accuracy for all datasets with at least one algorithm. Each algorithm was tested on randomized subsets of each dataset for 10 iterations. In all cases where there were more than 20 participants in the dataset, each iteration was done with a random choice of 20. Thus we can achieve very high accuracies if we carefully select our participants and algorithms.

	EEG	NBA Stat.	Act. Recogn.	Walking Act.	CAT Scan
Neural Network	0.51	0.95	0.82	0.5901	1.00
Random Forest	0.52	0.96	0.92	0.7119	1.00
SVM	0.3297	0.79	0.79	0.57	1.00

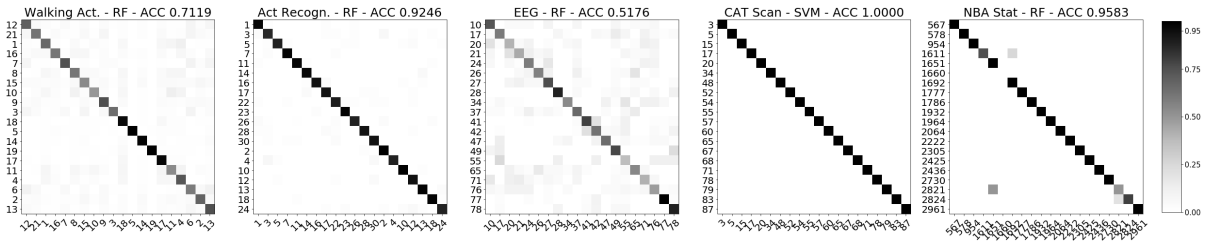

Fig. 3. Each confusion matrix and accuracy reported in this figure represents the best achieved performance across 10 iterations. A solid black square on the main diagonal means that the participant represented by this identifier is easy to identify. In all cases there were at least a few participants that were easy to identify which implies that their measurement values were distinct. The EEG dataset was the only case where the off-the-shelf algorithms were unable to achieve the accuracy goal across all participants. The axes contain the numeric identifiers used in the most favorable run, the identifiers have no ordering.

3.3.2 Confusion Matrix (CM). To illuminate the identification performance we computed the confusion matrix. It is a contingency table that tabulates how often one participant identifier is confused for another. The confusion matrix allows us to identify when participants fail to be distinct. Figure 3 shows both of the accuracy score and confusion matrix results.

3.3.3 Number Of Easily Identified Users. A user is easily identified if their measurements are classified correctly most of the time. This can be computed as a count of the main diagonal probabilities which are above a threshold (e.g. 80%). Table 4 enumerates the number of participants that can be easily identified by the classifier.

The strength of this metric lies in its dependence on how the measurement values separate participants. Because this metric is more sensitive to measurement separation than accuracy, it gives a more meaningful summary of the identification systems performance. Using this second metric we can see that the accuracy does not always give the full picture. For example in the NBA dataset, random forest has a slightly higher accuracy than neural networks, however, neural networks easily identifies more participants. This would indicate that neural networks are better at finding the structural separations in this type of measurement data. In contrast, random forests do significantly better at separating users in the walking activity data. Finally, support vector machines have a slight advantage in the CT scan data, because of the accuracy metric, even though all algorithms can easily classify all participants. *Multiple metrics help to identify cases where a single metric is artificially high because of artifacts in the data.*

In summary, we noticed the performance of identification systems varies across the different data sets and classification algorithms. We will analyze the reasons that cause the differences in performance.

Table 4. Easy Identification - We define participant easily identified when they are identified correctly at least 80% of the time. Note that the accuracy of random forest is higher than neural network in the NBA case (see Table 3), though neural networks easily identifies more participants. This is because accuracy only considers correct decisions without concern for whom they occur. In these datasets, random forests make more correct decisions overall, but neural networks have more certainty per individual.

	EEG	NBA Stat.	Act. Recogn.	Walking Act.	CAT Scan
Neural Network	1	18	12	2	20
Random Forest	1	16	20	5	20
SVM	1	12	12	1	20

4 THE IMPACT OF DATASET PROPERTIES ON CLASSIFICATION PERFORMANCE

The reliability of an identification system depends on the generalizability of classification. Generalizability here means how the classifier performs with participants which it was not trained on. To achieve such generalization, the dataset will need to be both representative of the intended user base, and contain enough information to ensure that the distributions governing the observations of measurements are approximated well. Predicting the generalizability of a system is often a difficult task because the properties of the reported datasets can only be used to gauge expected performance under certain constraints.

The representativeness of a dataset is related to the diversity of participants and dimensionality of the feature space. The expectation is that measuring more observables from a diverse array of participants would yield a model that has better coverage of the intended users. We show in this section that it is very difficult to define the diversity of a dataset. For example, we often lack ground truth about its intended user base. We also demonstrate that feature dimensionality rarely predicts performance because the size of the feature space is not a predictor of discernibility.

The size of the dataset is often used to gauge whether there are enough measurements to deem the approximation of a distribution sufficient. Generally, more datapoints yields a better approximation of a distribution [27]. However, it is often unclear which distribution is being approximated. We argue that sample size is poorly defined and neither of the potential definitions is sufficient to predict the error in approximation of the distributions that the system is attempting to learn. There are two types of distributions that impact performance and each definition of sample size is only related to one of them.

We compare impact of variation in the dataset properties, including measurement count, participant diversity and feature dimension (Table 2) on system performance when the participant count was fixed at 20. All of these properties of the data sets may affect performance because all algorithms attempt to optimize their fit of the data [1]. These differences between datasets can be lensed through the distributions that the identifier is trying to learn.

4.1 What is Sample Size: Participant Count or Measurement Count?

Sample size is a term used often, but unfortunately ambigious. Different disciplines do not agree on the definition of a sample [7]. Different disciplines tend to focus on different aspects of the analysis, the term sample size get used in two ways.

Often in the HCI literature, sample size means the number of participants. However, in the machine learning literature, sample size is usually used to refer to the count of measurements taken across all participants.

To avoid confusion, we will explicitly identify either the participant count, N, or the measurement count, T. One critical observation is that a large measurement count does not imply large participant count. The number

of participants within a study and how the participants were selected impacts sample diversity [11, 50], which is a gauge of how different the measurements from distinct participants are.

4.1.1 Why Do the Differences between Participant Count, N, and Measurement Count, T, Matter? The classification performance is controlled by two types of distributions: 1) the distribution of measurements taken from all participants, $P(V)$ (also known as the population distribution), and 2) the distributions of measurements from a single participant, $P_h(V)$ where h is a unique index for the participants. We will call this the individual distribution. Both of these distributions are approximated by machine learning algorithms when it fits a curve to the points in the dataset. The errors in approximation are directly related to the number of points within the dataset, but each of the different counts is only related to a specific distribution. We argue that participant count, N, can be used to gauge the error in approximation of the distributions among participants if there is sufficient participant diversity. We also argue that measurement count, T, alone is insufficient to gauge whether either distribution is well-approximated because it does not represent how many measurement points exist per individual. Thus, a critical difference between N and T is that, under specific conditions, N, can be used to compare the evaluations done between two systems, while T cannot.

The population distribution, $P(V)$, is illustrated in Figure 4 (B) and (D). It captures how a specific measured observable (or sets of measured observables) varies across the intended user base. It is approximated by the relative frequency of a measurement values collected across a sampling of that user base. The individual distributions, $P_h(V)$, are illustrated in Figure 4 (A) and (C). Since the measurements from each participant may have different variation characteristics, each participant, h, has their own distribution $P_h(V)$. How these $P_h(V)$ distributions overlap will impact how well the measurements separate, and ultimately dictate the system performance.

If there is sufficient participant diversity, then larger participant count, N, may imply a better approximation of the population distribution, $P(V)$. In the absence of a systematic selection biases, as N increases, so do the chances of observing distinct measurement values. Thus, when N is sufficiently large, we can consider $P(V)$ well approximated as more of the range of measurement values has been explored.

Total measurement count, T, is insufficient to ensure that $P(V)$ is well-approximated. The number of measurement per participant, M, also impacts the approximation. If we assume that each individual distribution, $P_h(V)$, is well-approximated and that the sampling has no systematic bias, then a larger N may also imply that the set of distributions has better coverage over the range of possible measurement values.

Both distributions are necessary to estimate the details of data. The population distribution, $P(V)$, can inform the sampling procedure. If we knew the population distribution, we could build our recruiting policy using standard techniques [4, 8, 9, 29, 32, 44], allowing us to answer questions such as "For a given measurement value (or values), V, how likely, $P(V)$, is it to observe this V from the population?" However, the population distribution cannot discern whether the measurement type effectively separates individuals because the distribution lacks that granularity. By looking at the overlap in individual distributions, $P_h(V)$, can be used to gauge if the measurements are adequate for distinguishing individuals in the sample of the intended user base. However, the $P_h(V)$ distributions can not discern if there are additional distinct individuals among the intended users that were not accounted for. To answer this question we could use the population distribution, $P(V)$, to identify probable measurement values that are not present in our data (e.g. values beyond 5 and -5 in Figure 4). The individual distributions, $P_h(V)$, also cannot assess the likelihood that an unseen user is easily confused with the members of the sample. To solve this issue, we would need to compute the conditional probabilities of overlap which uses the $P(V)$ distribution as a prior. Thus, a purely mathematical analysis would require both distributions in order to determine how many individuals a system must be tested with before the performance degrades below tolerable levels.

Given that distributions of users are rarely known, user studies help to estimate these distributions. In most systems, there is an enrollment phase [48] in which several measurements are taken from all the participants,

h, that are available and an estimate of $P_h(V)$ is made. This process is applied to each h which may use the system. When the system encounters new, unlabeled measurements, it makes a decision using these estimated distributions. The quality of the distribution estimates is a function of the number of measurements taken from each individual. Since each individual has a different level of variability for any type of measurement, the number of measurements required to get a good estimate will generally not be the same across individuals. Even though these differences are present, in practice a large number, M, is selected which accommodates some range of variability across all h and ensure that all $P_h(V)$ are well-approximated. Under this assumption $M * N = T$.

For any given measurement value, V, if only a small number of participants have a high probability of producing that value, the performance will be high. This is a property of what is being measured, and we consider the measurement discerning when this happens (case (A) in Figure 4). When this is true, we may be able to identify individuals in the sample population easily. In contrast, if many individuals produce the same V when measured, then the probability, $P_{h'}(V)$, of observing a V from a randomly chosen individual, h', is high, thus this measurement type will be less useful for identifying individuals (case (C) in Figure 4)). The multiclass classification problem tries to identify h for an arbitrary V by considering all the probabilities, $P_h(V)$, across all h for which we have an approximation of the individual distribution. In the simplest case we can ask the question "for any V which h was the most likely to produce this V?" (however the decision logic is often more complex, e.g. taking into account correlations between individuals).

4.1.2 The Impact of M and N on Performance. It is difficult to gauge how effective a system will be at identifying individuals just by looking at the number N of users used to test it. Consider the examples of identifiers built for the EEG dataset from section A.4 and the CT scan dataset of section A.3. We observe significantly different performance (CT scan $ACC \approx 100\%$ vs. EEG $ACC \approx 52\%$) for two datasets where the number of participants used to build the identifiers was kept the same, $N = 20$. For each dataset the average number of measurements per participant, M, was comparable (see Table 2). Therefore, the total number of measurements, $T \approx M * N$, was comparable. The significant difference in performance can only be explained by the discernibility of the feature space, which is a function of the individual distributions, $P_h(V)$. The performance of the random forest classifier on the CT scan dataset only dropped by $\approx 1\%$ when the participant count was increased to $N = 80$ (See Figure 5). In contrast, using the random forest classifier on a subset of 10 participants for the EEG dataset achieved an increased accuracy of $\approx 65\%$.

A system with more samples per participant (higher M) will not necessarily yield better results. For example, the NBA data set of Section A.5 has an order of magnitude fewer measurements per participants than the activity recognition data set of Section A.2, yet it achieves $\approx 3\%$ higher maximum accuracy. If the feature space is highly discerning, then under-sampling the participants may not cause significant degradation in performance because the individual distributions are spread apart. On the other hand, if a feature space is not discerning, sampling each participant further will not produce any improvement. The approximations of the individual distributions will become tighter, but distribution overlaps will remain the same.

4.2 How Participant Diversity Affects Performance

The two types of distributions from Section 4.1 highlight a key challenge when formulating a participant recruiting policy: ensuring that you have covered the breadth of measurement variation within a population. To be sufficiently representative, a dataset must collect measurements from a wide range of distinct individuals in order to determine if we have adequately covered the range of measurement values that have non-trivial probabilities in the population distribution, $P(V)$. If the value range is not covered, the generalization beyond the participants recruited will suffer because a system built from this dataset will encounter measurement values in the intended users that are significantly different than the values with which it has been trained. These circumstances render the behavior of the system indeterminate.

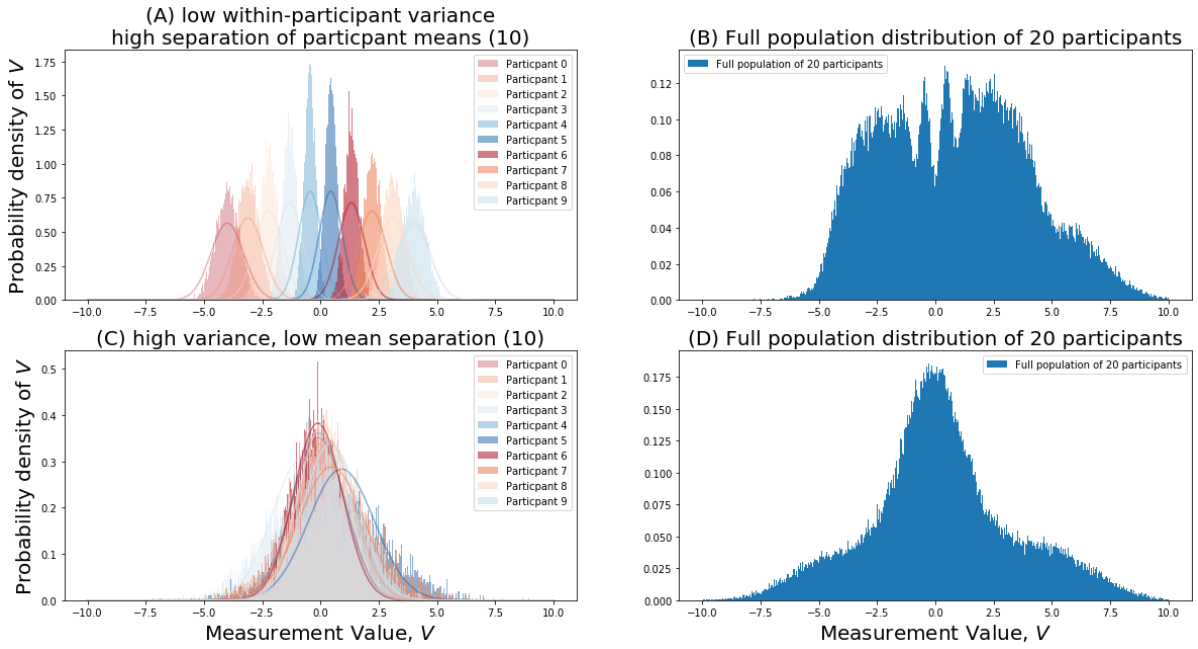

Fig. 4. The variation of a measurement's value has two possible sources. The first is random fluctuations in the value when repeated measurement are taken from an individual (i.e variation within an individual). The second is random fluctuations in the value when measurements are taken from many individuals through out the population. If the variation within an individual is small compared to the variation among individuals then these measurements may separate participants well. For example, the ten participants in (A) separate well. When the variation condition holds, any measurement value, V_h from a particular participant h, has a low probability, $P_{h'}(V_h)$, of coming from another participant h'. For example, repeated CT scans within a small time window of a single individual should have very little variation, however, these scans should be very different between individuals because they measure the entire body. In contrast, when the variation among participants is high compared to the variation among users (C), all measurement values have comparable probabilities of coming from each user (e.g. EEG data has so much noise in the measurement that any individual value can easily have come from many individuals). The population distributions shown in (B) and (D) show the measurement range of the full population of 20 participants. This distribution can tell us if a measurement value is reasonable for a population, that is, has a non-zero probability of being observed. It cannot be used to determine whether the measurements are useful for distinguishing individuals.

For any population, diversity refers to the degree of difference between members of that population. Other disciplines, such as ecology, attempt to categorize the variation within in a population by computing the Shannon entropy of the probability distribution on specimen observation [11]. Such metrics often assume that all specimens are readily distinguishable from each other, that is, low variation within an individual. It is expected that when a rare specimen is observed, the observer would be able to easily recognize that the observed specimen is distinct from the previously observed samples.

In comparison, it is more difficult to determine population diversity within the context of identification systems, because we are viewing each participant through the lens of limited precision measurements. In Figure 4 we observed cases where the participant count, N, was inadequate to cover the entire range of possible measurement values, V. A fully characterized feature space requires the recruitment of more participants. If the measurement

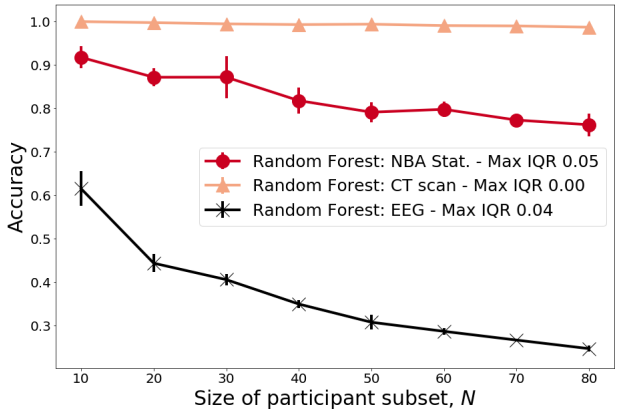

Fig. 5. We can compare the accuracy of the random forest classifier when run on three of the data sets with enough participants to increase N significantly. For each N, we perform 10 runs with randomized subsets of size N. The CT scan dataset has significantly less variation in the performance metric. We measured this by looking at the interquartile range of performance metric values when the given randomized participant subsets. This variation in the CT scan dataset was an order of magnitude lower that the other datasets. The range is not visible in this figure because of scale. Because the CT scan dataset is resilient to variations across humans, its performance degrades much more slowly as the number of participants increases. The variability between randomized subsets may give an indication of how performance will degrade as the participant count increases.

values not observed are rare (e.g. 7.5 in case (B) of Figure 4), then we may need to recruit significantly more participants before we observe these rare values. If we could sample until we covered the full range of values, we would ensure that we have a reasonable participant diversity. Unfortunately in most cases, the range of values is not known apriori.

The number of participants in a study, N, cannot always be used to directly determine the if the study was sufficiently diverse because of the break down in distributions covered in Section 4.1.2. Consider the participant counts of the CT scan and the EEG datasets in Table 2). The two datasets have similar number of participants, but the best achieved performance of an identification system built from EEG dataset was 65% accuracy when the participant count was 10 (see Figure 5). As the participant count increased, the system performance degraded. We argue that the system is well tested because we can identify a participant count beyond which the performance guarantees no longer hold. In contrast, the performance of the identification system built with the CT scan data does not degrade significantly as the participant count increases. For this identifier, the user limit is unknown because our tests could not produce a reduction in performance even with 80 users. This difference in performance as the participant count increases will play a large role in the robustness of systems built upon these identification systems.

4.3 How Feature Dimension Affects Performance

Feature dimension is a count of the number of distinct measurement types and functions of the measurements values that make up feature vector. It does not consider redundancy among features in the feature space nor does it contain any information about the distribution of measurement values. For example, if we use a length as a feature, this length could be measured in meters or kilometers. Since one value is simply a scalar multiple of the other, good algorithms will treat them as the same feature. Techniques such principal component analysis can

be used to reduce the feature dimension by considering the minimal number of vector components required to represent the information within the feature space.

Feature dimension is rarely useful for predicting how well a system will preform. Despite this, it is often reported. It cannot be used to perform relative comparisons of identification systems. For example, the identifier built from the walking activity data set uses only four features (see Table 2), but still achieves ≈ 71% accuracy (see Table 3). In contrast, the EEG data set identifier has 39 features to work with but performs significantly worse (≈ 51% at best) than the walking activity data set identifier. Further still, the NBA data set identifier uses 51 features and achieves ≈ 95% accuracy. The CT scan dataset identifier has a lower feature dimension to work with than the activity recognition dataset identifier, but performs better. The ability of a system to distinguish individuals is only as good as the discernibility of measurements will allow.

5 BOUNDS ON THE NUMBER OF EASILY IDENTIFIED PARTICIPANTS

All identification systems degrade in performance as the number of users increases. This is because the system reaches the upper bound (or the upper limit) of easily identifiable users. In this section, we show that the upper bound arises because each measurement obeys a natural distribution across the intended users, $\mathcal{P}(V)$.

The natural distribution, $\mathcal{P}(V)$, can only assign non-zero probabilities to a finite range of values. All participant measurements values are drawn from this range. As the participant count, N, increases, it becomes increasingly probable that any measurement value will be observed from multiple participants. To be able to predict when the system will fail, we need to know at what participant count, N, performance begins to degrade beyond tolerable levels. The upper bounds in all systems differ in how large that bound is, and how quickly it can be found in a practical setting. In the example of the identifier built from CT scan dataset of Appendix A.3, the upper bound is very large (see Figure 5). Thus, any system based on these measured observables would be able to distinguish a large number of participants. It would not be practical to build an identification system that requires a full body scan of an individual to perform the identification. However, the system demonstrates a case where machine learning algorithms do most of the work with minimal configuration.

We can notice trends in performance with a scatter plot of the principal components as N increases. Figure 6 shows four scatter plots of the first two principal components for the NBA stats dataset. Each scatter plot is labeled with the achieved accuracy and participant count, N. When participant count is small ($N = 20$), the measurement separation is very good, and thus, the classifier will easily discern participants. As N increases, we can see crowding within the center of the graph. This crowding is a form of concept drift [20], where the conditional distribution of a participants identifier given the measurement, $P(h|V)$, changes as the number of participants grows. In our case, an increased participant count is the source of the distribution shift as opposed to temporal drifts which are normally observed. This comparison demonstrates that performance analysis with a small number of participants (compared to the bound on participants that can be represented) is incomplete at best.

5.1 Why Are the Metric Values of Low Participant Count Studies Misleading?

If variability of measurements within an individual is low enough to produce distinct participant measurements, such as those shown in Figure 4 (A), then the system will only make mistakes if the recruiting procedure produces two individuals with the similar individual distributions, $P_h(V)$ (e.g. similar location and scale). In a Bayesian formulation, the equivalent condition is $P(V) \propto P(h)$.Thus, the performance of either class of identification system with small number of distinct participants, N, is largely dictated by the population distribution $P(V)$. The distinctiveness of participants implies that the probability of observing a specific measurement value is very dependent on recruiting a specific individual.

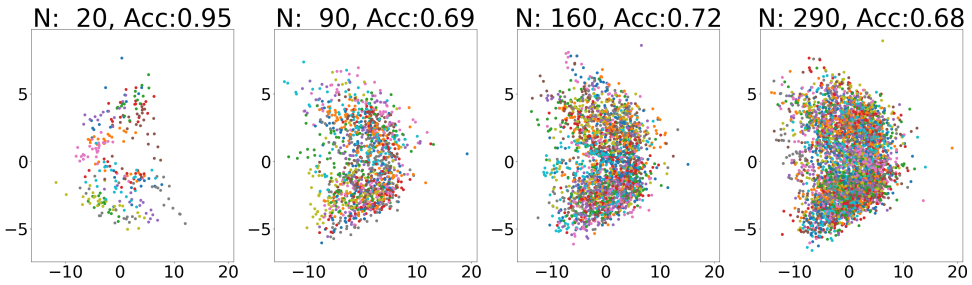

Fig. 6. As we increase the size of the participant count, N, the first two principal components become more crowded. The feature space in often multi-dimensional and these two principal components are the most variable linear combinations of the components in the feature space [18]. The clustering of points in this figure is a two dimensional representation of the per-individual distributions, $P_h(V)$ discussed in Figure 4. This crowding explains why the performance of the identifier will degrade with increased N. When N is large, the various point distributions overlap into an incoherent mass.

The identification system can be thought of as partitioning the space of measurement values into bins which correspond to the classes (as in Figure 7). In the case of user identification, each distinct participant's measurements might neatly fall into a bin designated for them. As long as the participant count is low, the values in each bin will mostly have come from distinct participants with overlaps being rare. Such a condition would artificially inflate summary metrics that try to count mistakes, for example, accuracy or the confusion matrix, because mistakes are artificially rare. As the number of participants increases, $N \nearrow$, it becomes more likely that a bin will have measurements from more than one participant. If there are naturally $\mathcal{N}$ distinct bins, then the pigeonhole principle [42] guarantees that there will eventually be at least one bin with multiple measurements from different participants and thus the system will start to accumulate errors which will drive accuracy down.

For multiclass identification systems, the number of bins may not coincide with the number of participants. In many cases the raw measurements are mapped into a different representation via a deterministic function to achieve the same bin separation. When the number of bins is smaller than N, a single participant might get mapped into multiple bins. Although this mapping may lower the effective number of participants required by concentrating them into fewer classes, it does not eliminate the problem that too few participants underexplores the space of possible inputs. Thus, it is still necessary to test with increased participant count to ensure that the space of possible measurement values does not contain values that fail to be mapped.

Machine learning algorithms are designed to optimize the amount of information that is extracted from measurement data. This optimization tries to bin measurements to have a maximal separation between classes. This separation translates to classification performance, for example, accuracy. The problem of misleading metrics arises when the participant count is so small that the ability to draw these boundaries is artificially easy. When this occurs, *the performance of a system is less influenced by the effectiveness of the measurements at separating participants and more influenced by the how diverse a sample population the recruiting process produces.*

6 AN ITERATIVE APPROACH TO TESTING SYSTEMS

In this section, we describe an iterative approach for testing identification systems. Instead of setting a goal number for participants N in the beginning of the study, we can keep increasing N while studying the system and its performance. This is in contrast to statistical group comparisons in experimental designs using null hypothesis statistical testing (NHST) where you must set the target N in advance.

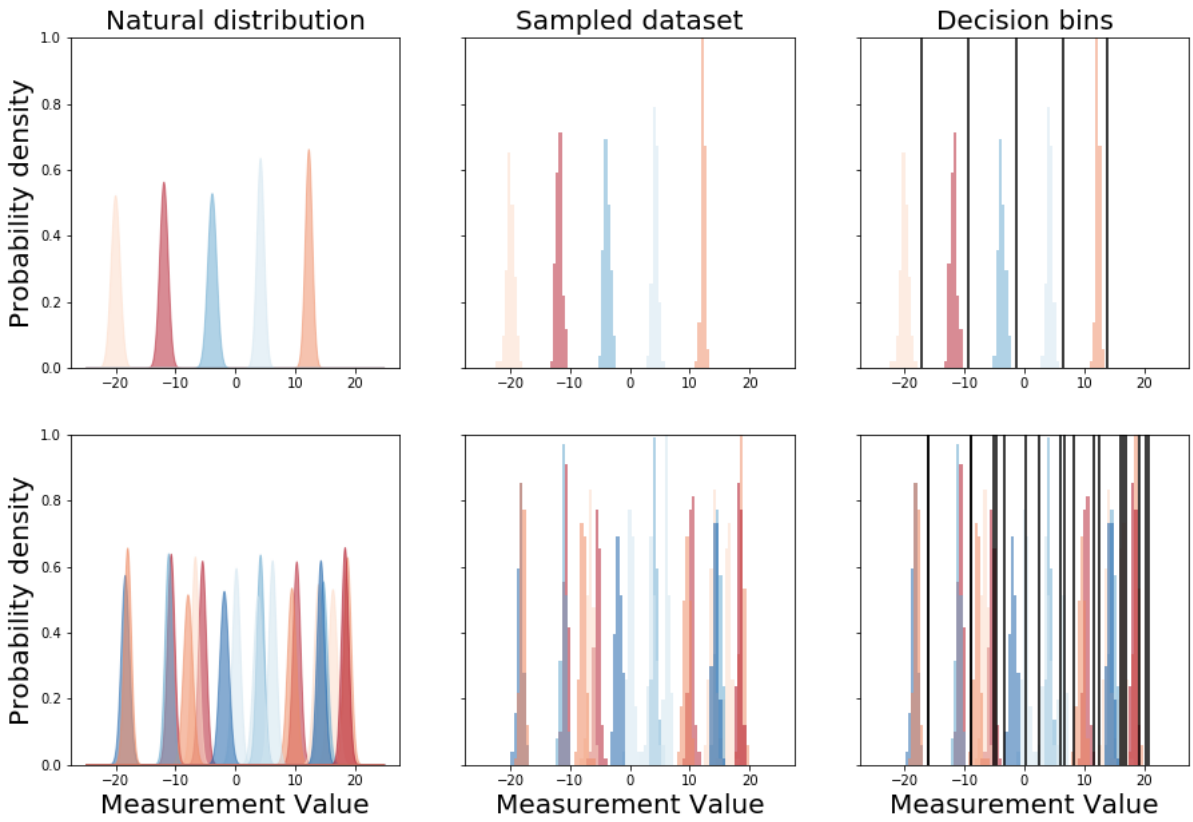

Fig. 7. When the variation within participants is low (low variance in $P_h(V)$) and the participant count, N, is small, the probability of observing a measurement value is proportional to the probability of picking that participant, $P(V) \propto P(h)$. At low N, samples drawn from the natural individual distribution, $\mathcal{P}_h(V)$, separate well. Thus, it is very easy to draw the boundaries of the bins (e.g. the top row where $N = 5$). As the number of participants increases, the $\mathcal{P}_h(V)$ over lap, thus the samples will overlap, and the bins become more difficult to draw (e.g. the bottom row where $N = 20$). The identification system tries to optimize the placement of bin boundaries by using the sampled dataset as an approximation for the natural distribution.

We cannot make performance guarantees on identifications system by only knowing the number of participants N. This is because we do not know the upper bounds as described in Section 5. Although it may not be possible to know how far we are from the bounds, we can gauge how the performance degrades as N increases. This can serve as a method for relative comparison of systems. We can take an iterative approach where we assess how the performance metrics of the system react to increasing N without starting a study with a large set of participants. Instead, we can iteratively add participants to the study until we have identified an N that causes the performance to degrade below a tolerable level. We describe this in Algorithm 1.

The analysis in Section 4 demonstrates no single property of the dataset is a good indicator that the resulting system will perform well. Individually, none of these values can guarantee that the experiments conducted truly tested the generalization limits of an identification system. Even when these values are optimal, such systems might still be susceptible to unexpected identification errors if participants were chosen with some systematic bias

which causes the dataset fail to be representative. It is usually not possible to know the variability of the measured quantities apriori. Thus, it is difficult to construct a practical recruiting policy that eliminates all possible bias before collecting some measurements.

By establishing trends in how the performance varies when different subgroups of participants are selected and degrades when the sample diversity increases, we can determine if a system requires further testing. Cycling the participants into randomized subsets compliments the approach of increasing participants. It can identify subsets of the population that are artificially distinctive, which would produce higher than normal accuracies. It can also identify subsets which are very similar, which would lead to lower than normal accuracies. This cycling provides another check on how brittle the models learned from a specific size subset are. Sample diversity itself is difficult to measure directly. Instead, we can use participant count, N, as a proxy for sample diversity with some considerations. We need to ensure a reasonably sized number of samples per participant, M. This M, will can be chosen as the largest number of samples required to ensure good approximation of $P_h(V)$ for all h. It has to be determined empirically after the initial set of measurements is taken by looking at the variance of the distribution estimations. We would also need to eliminate systematic participant selection bias in our recruiting process by identifying factors in the process that might limit the range of measurement values.

Algorithm 1: Iterative approach to testing

```
Data: Initial n
Result: Plot of performance metric vs n
Collect m labeled measurements from all n participants (where m is sufficiently large as in Section 4.1.1);
Choose randomized subsets from n participants;
foreach Randomized subset do
    Build model on data of subset;
    Compute performance metric for each model;
end
Compute interquartile range (IQR) as a measure variability of the performance metric across subsets;
while Performance metric above tolerable level and performance metric unstable do
    Increase number of participants to n' ;
    Collect m labeled measurements from new participants ;
    Choose randomized subsets from n' participants;
    foreach Randomized subset do
        Build model on data of subset;
        Compute performance metric for each model;
    end
    Compute IQR;
    n ⟵ n';
end
Plot metric against n with interquartile range error bars;
```

6.1 Test with Increasing Participant Counts, N

The rate of degradation of the performance metrics is both a function of the dataset and the machine learning algorithms. We argued in Section 5 that performance degradation as participant count, N, increases will occur regardless of algorithm or sampled dataset. However, each system will differ in the rate of degradation because each

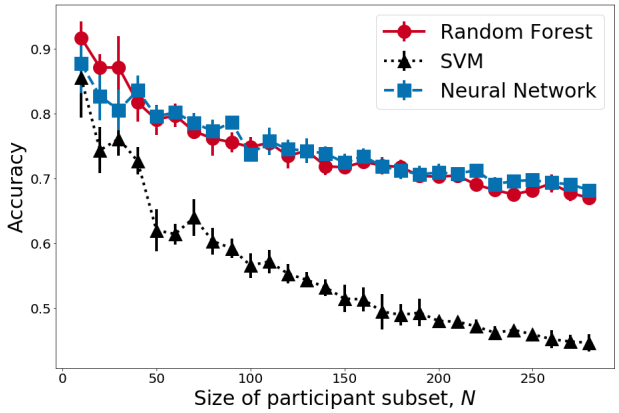

Fig. 8. In the NBA dataset, we can observe a drop in accuracy as the participant count, N, increases regardless of which algorithm was used. For each algorithm, we trained a model with N participants taken from the full population of 290+ players. We then calculated the accuracy for each model and repeated the process for ten iterations. The graph shows the median value and the interquartile range used as error bars. As N increases, we can see that all models' performance degrades. However, some are more impacted than others. As we noted in Section 3.2, each algorithms performs differently when given different types of measurement data. Although these differences indicate that the degradation rates will not be the same, it should be noted that eventually all algorithms degrade when participant count is large enough.

algorithm has a different efficiency for extracting information from the dataset and each dataset's representation of the natural phenomenon begin measured is of varying quality. The differences in the rate of degradation can be used to compare both algorithms (see Figure 8) and measurement types (see Figure 5) as each exhibits differing rates of degradation as N increases.

The rates of degradation may differ for each algorithm and dataset pair. The bound on easily identified users is largely a function of the measurement type. Specifically, it depends on the amount of information that can be extracted from a measurement type. We will always observe a steady degradation in performance as new participants are added, even in the ideal case. The identification system will only experience gradual drops in performance. For example, it not possible that the accuracy remains constant at 100% up to $\mathcal{N}$ participants and then goes to 0% with the $\mathcal{N}+1$ participants, unless all $\mathcal{N}$ participants are replaced with a new set of $\mathcal{N}$ participants that are distributed differently. Knowing the rate at which new participants degrade the performance gives us an idea of where the bound on users that can be identified while satisfying a constraint on error occurs. The perfect identifier scenario would lead to a degradation rate that is proportional to the probability of observing each individual in the population (see Figure 1). However, most identifiers will not be perfect, and the rate of degradation will often be faster than this ideal scenario. In Figure 5, we observe that a single algorithm (parameterized in the same way) has very different performance degradations as the number of participants, N, increases up to 80. The various sensitivities to increasing participant count gives us an idea of how the system will fail when the participant count grows larger than anticipated.

6.2 Test with Randomizing Subsets

When the number of participants is limited, an additional check for unstable identification performance is to select multiple randomized subsets of participants and evaluate models derived from these subsets independently. Similar to leaving-out-k cross validation [18], we obtain multiple values for the performance of the system. Randomizing the subsets differs from standard cross validation in several ways. First, when a subset of participants is chosen,

the data is partitioned into a training and testing portions. The model is trained and tested with the respective portions. This includes computing the performance metric. Participants not chosen for this subset will not be evaluated against this model since no training data was present for them. Additional analysis could be done by testing the system with these left out members, though this analysis would produce a different performance evaluation (a test of out-experiment generalization). Secondly, the distributions being trained on changes with every subset, instead of being drawn from a common pool as is the case in cross validation procedures. Each subset iteration builds a new model which reflects the current subset. Thus, the process of randomizing subsets is not testing a specific model, but instead the ability to construct discerning models across a breadth of user sub populations.

Figure 5 shows how different measurement types exhibit significant variation in scale. This variation is a function of the discernibility of measurement type. If a measurement type has high discernibility, then the performance metrics will be stable across a wide range of participants.

By cycling different participant subsets, we increase the likelihood that our system will encounter subsets of participants that may artificially increase performance. When the algorithm is fixed, each subset will yield different overlap behavior for natural distributions, $\mathcal{P}_h(V)$, for all h in the subset. These distributions will result in differing boundary choices (see Figure 7). The differences between subsets will ultimately dictate the variation in performance. If the variation in performance across subsets is significant, then the performance may degrade quickly as the participant count, N, increases because participants that are hard to classify will be added to the subset. On the other hand, if the variation across subsets is not large, we will need to test with more participants to determine the limit on easily identifiable participants.

We can measure the spread of performance metric values (e.g. IQR or variance) as the subsets are varied to quantify how much a metric value can change when a subset is favorable. If the subsets are very different, the spread will be large. In Figure 5, for each N, ten participant subsets of size N were chosen from the larger dataset. Then, a classifier was trained and tested with the data from these ten subsets. The CT scan identifier barely degrades as N increases, and the error bars for the metric IQR across all subsets are so small that it is not visible on the plot. In contrast, the identifier built on the EEG dataset starts with varying performance, but this variation stabilizes as the performance degrades. The performance stabilizes because the subset distributions become more stable (and overlap significantly more) as the participant counts increases. Thus, after some intermediate N, it becomes more difficult to find a subset of the population which will separate well by chance. Since the performance metric variability on subsets is only a weak gauge of the potential performance degradation, the technique of cycling subsets of participants should be used to augment the analysis with increasing participant counts.

7 RELATED WORK

The subject of sample size has been discussed for decades [29, 44]. Studies have determined sample size with power analysis where the shape of the distribution is assumed to belong to a specific family [30, 32]. In these cases samples size means participant count and the number of measurements per participant was one. This class of analysis is not applicable to studies where the goal is to build a decision system based on multiple measurements from many diverse participants. The assumed distribution families are too simplistic and the number of measurements from a single participant is often greater than one.

The issue of low participant count in user studies is common across different research communities. Caine [9] analyzed the sample sizes of all 465 manuscripts in the proceedings of CHI 2014 and found the common sample size is only 12. In the HCI community, researchers may try to mitigate the issue by collecting more data from the same participants. However, we showed that this approach is not adequate because each participant can only provide a limited amount of variation for a given measured observable.

Many other fields have raised concerns when the participant count is low. Raudys and Jain [40] discussed the influences of sample sizes on feature selection and error estimation for different types of simple classifiers such as Euclidean distance classifier and Fisher's linear discriminant. Button et al. [8] showed the average statistical power of studies in neurosciences is very low. They emphasize that this situation leads to overestimates of effect size and low reproducibility of results. Anderson and Vingrys [4] proposed three situations that need to be considered while conducting research with small sample sizes in psychophysical and neurophysiological studies. Hackshaw [25] overviews the strengths and limitations of small sample size in clinical studies. The main issue for small participant count studies is that the outcomes have large standard error and no firm conclusions.

Our survey shows that small participant counts are an issue in user and multiclass identification studies. The breadth of previous work focused on trying to learn population distribution parameters of an assumed distribution shape. These shape assumptions were used as a guiding principle to select participant count. Bounds on error were established only as an afterthought by considering measurement variation after the procedure was done.

In our approach, we do not make strong assumptions about the shape of the population distribution to determine when our statistics have converged. Instead, we apply an iterative approach that uses the error metrics to identify the limits of the system's ability to discern individuals. We update our model as new data is made available and estimate distributions we cannot know a priori. Our approach provides a method for reasoning about a problem that is often poorly defined without making assumptions that limit generalizability. Our approach is adaptive and can react to shifts in the population and unexpected experimental conditions.

8 DISCUSSION AND CONCLUSIONS

We have shown that testing an identification system with a small number of users is rarely adequate and often misleading. To demonstrate this, we constructed five user identification systems from publicly available datasets. Three of these systems yielded $\geq 90\%$ accuracy when the participant count was small.

To explain why such misleading results can arise from low participant count user studies, we delved into the properties of the measurements that would impact these metrics. We demonstrated that as the participant count increases, the system performance must decrease. We reasoned that because all measurements can only be made with finite precision, an upper bound on the number of easily identifiable individuals must exist. As the participant count approaches this bound, the performance of the system must degrade.

We showed the issue of low participant count is common in the user and multiclass identification studies by surveying the recently published papers in top-tier venues. Seventy-seven percent of the 30 surveyed papers were supported by user studies with 20 or less participants. Although some of the work collected thousands of measurements from the small participant sets, we argued that these measurements do not compensate for a lack of sample diversity, which is mainly affected by the variation between participants.

We showed that the participant count can be used as a proxy for sample diversity, given that the user study factors are controlled. We can establish an estimate that gauges how the system performance will degrade when the participant count increases. We demonstrated that performance metric variation on randomized participant subsets can be a useful approach to diagnose performance degradation when the participant count increases. Knowing these factors will enable us to reason about the likelihood of failure for a system in a target application.

To conclude, we argue that limit on easily identified participants can and should be experimentally determined by increasing the participant count iteratively. There is no single participant count that will be sufficient for every experiment. As such, we cannot prescribe a fixed value or gauge what a large value would be. To learn how a system performance degrades when the number of participants increases, it is critical that we recruit until it fails.

REFERENCES

[1] Yaser Said Abu-Mostafa. 2012. *Learning From Data* (1st ed.). AMLbook.com, Pasadena, CA, Chapter Training versus Testing, 39–69.

[2] Amazon. 2019. http://amazon.com/go.

[3] Amazon. 2019. https://www.amazon.com/.

[4] Andrew John Anderson and Algis Jonas Vingrys. 2001. Small Samples: Does Size Matter? *Investigative Ophthalmology and Visual Science* 42, 7 (2001), 1411.

[5] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649

[6] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.. In *ESANN*.

[7] Peter Bruce and Andrew Bruce. 2017. *Practical Statistics for Data Scientists: 50 Essential Concepts*. " O'Reilly Media, Inc.".

[8] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 5 (2013), 365.

[9] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[10] Pierluigi Casale, Oriol Pujol, and Petia Radeva. 2012. Personalization and user verification in wearable systems using biometric walking patterns. *Personal and Ubiquitous Computing* 16, 5 (2012), 563–580.

[11] Anne Chao and Tsung-Jen Shen. 2003. Nonparametric estimation of ShannonâĂŹs index of diversity when there are unseen species in sample. *Environmental and ecological statistics* 10, 4 (2003), 429–443.

[12] Xiang 'Anthony' Chen and Yang Li. 2016. Bootstrapping User-Defined Body Tapping Recognition with Offline-Learned Probabilistic Representation. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 359–364. https://doi.org/10.1145/2984511.2984541

[13] Yuanying Chen, Wei Dong, Yi Gao, Xue Liu, and Tao Gu. 2017. Rapid: A Multimodal and Device-free Approach Using Noise Estimation for Robust Person Identification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 41 (Sept. 2017), 27 pages. https://doi.org/10.1145/3130906

[14] Sauvik Das, Gierad Laput, Chris Harrison, and Jason I. Hong. 2017. Thumprint: Socially-Inclusive Local Group Authentication Through Shared Secret Knocks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3764–3774. https://doi.org/10.1145/3025453.3025991

[15] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[16] FaceBook. 2019. https://www.facebook.com/business/ads.

[17] Judith M Ford, Vanessa A Palzes, Brian J Roach, and Daniel H Mathalon. 2013. Did I do that? Abnormal predictive processes in schizophrenia when button pressing to deliver a tone. *Schizophrenia bulletin* 40, 4 (2013), 804–812.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, NY, USA:.

[19] Matteo Gadaleta and Michele Rossi. 2018. IDNet: Smartphone-based gait recognition with convolutional neural networks. *Pattern Recognition* 74 (2018), 25 – 37. https://doi.org/10.1016/j.patcog.2017.09.005

[20] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.* 46, 4, Article 44 (March 2014), 37 pages. https://doi.org/10.1145/2523813

[21] Omri Goldstein. 2017. NBA Players stats since 1950. https://www.kaggle.com/drgilermo/nba-players-stats/home

[22] Jun Gong, Yang Zhang, Xia Zhou, and Xing-Dong Yang. 2017. Pyro: Thumb-Tip Gesture Recognition Using Pyroelectric Infrared Sensing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 553–563. https://doi.org/10.1145/3126594.3126615

[23] Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2011. 2D image registration in CT images using radial image descriptors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 607–614.

[24] Tobias Grosse-Puppendahl, Xavier Dellangnol, Christian Hatzfeld, Biying Fu, Mario Kupnik, Arjan Kuijper, Matthias R. Hastall, James Scott, and Marco Gruteser. 2016. Platypus: Indoor Localization and Identification Through Sensing of Electric Potential Changes in Human Bodies. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '16)*. ACM, New York, NY, USA, 17–30. https://doi.org/10.1145/2906388.2906402

[25] A. Hackshaw. 2008. Small studies: strengths and limitations. *European Respiratory Journal* 32, 5 (2008), 1141–1143. https://doi.org/10.1183/09031936.00136408 arXiv:http://erj.ersjournals.com/content/32/5/1141.full.pdf

[26] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-Id: Automatic User Identification and Authentication Through the Motions Captured from Sequential Human Activities Using RFID. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 48

(June 2019), 26 pages. https://doi.org/10.1145/3328919
[27] Steven M Kay. 1998. *Fundamentals of statistical signal processing: Detection theory, vol. 2.* Prentice Hall Upper Saddle River, NJ, USA:, Chapter 3, 61–65.
[28] Frederic Kerber, Michael Puhl, and Antonio Krüger. 2017. User-independent Real-time Hand Gesture Recognition Based on Surface Electromyography. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, Article 36, 7 pages. https://doi.org/10.1145/3098279.3098553
[29] Robert V Krejcie and Daryle W Morgan. 1970. Determining sample size for research activities. *Educational and psychological measurement* 30, 3 (1970), 607–610.
[30] John M Lachin. 1981. Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials* 2, 2 (1981), 93–113.
[31] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign Language Recognition Using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 23 (March 2018), 21 pages. https://doi.org/10.1145/3191755
[32] Robert C MacCallum, Michael W Browne, and Hazuki M Sugawara. 1996. Power analysis and determination of sample size for covariance structure modeling. *Psychological methods* 1, 2 (1996), 130.
[33] Jess McIntosh, Asier Marzo, and Mike Fraser. 2017. SensIR: Detecting Hand Gestures with a Wearable Bracelet Using Infrared Transmission and Reflection. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 593–597. https://doi.org/10.1145/3126594.3126604
[34] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807
[35] N/A. 2018. Data.gov. https://www.data.gov
[36] Netflix. 2019. https://help.netflix.com/en/node/100639.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[38] Ken Pfeuffer, Matthias J. Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 110, 12 pages. https://doi.org/10.1145/3290605.3300340
[39] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring Motion Direction Using Commodity Wi-Fi for Interactive Exergames. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1961–1972. https://doi.org/10.1145/3025453.3025678
[40] S. J. Raudys and A. K. Jain. 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 3 (March 1991), 252–264. https://doi.org/10.1109/34.75512
[41] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–59.
[42] Benoît Rittaud and Albrecht Heeffer. 2014. The pigeonhole principle, two centuries before Dirichlet. *The Mathematical Intelligencer* 36, 2 (2014), 27–29.
[43] Wenjie Ruan, Quan Z. Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: Enabling Fine-grained Hand Gesture Detection by Decoding Echo Signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 474–485. https://doi.org/10.1145/2971648.2971736
[44] Margarete Sandelowski. 1995. Sample size in qualitative research. *Research in nursing & health* 18, 2 (1995), 179–183.
[45] Munehiko Sato, Rohan S. Puri, Alex Olwal, Yosuke Ushigome, Lukas Franciszkiewicz, Deepak Chandra, Ivan Poupyrev, and Ramesh Raskar. 2017. Zensei: Embedded, Multi-electrode Bioimpedance Sensing for Implicit, Ubiquitous User Recognition. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3972–3985. https://doi.org/10.1145/3025453.3025536
[46] Stefan Schneegass, Youssef Oualil, and Andreas Bulling. 2016. SkullConduct: Biometric User Identification on Eyewear Computers Using Bone Conduction Through the Skull. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1379–1384. https://doi.org/10.1145/2858036.2858152
[47] Maximilian Schrapel, Max-Ludwig Stadler, and Michael Rohs. 2018. Pentelligence: Combining Pen Tip Motion and Writing Sounds for Handwritten Digit Recognition. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 131, 11 pages. https://doi.org/10.1145/3173574.3173705
[48] Michael E. Schuckers. 2010. *Computational Methods in Biometric Authentication.* Springer London. https://doi.org/10.1007/978-1-84996-202-5
[49] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 134 (Sept. 2018), 27 pages. https://doi.org/10.1145/3264944
[50] Edward H Simpson. 1949. Measurement of diversity. *Nature* 163, 4148 (1949), 688.

[51] Spotify. 2019. https://www.spotify.com/.
[52] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 189 (Dec. 2018), 27 pages. https://doi.org/10.1145/3287067
[53] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. 2019. Robust Performance Metrics for Authentication Systems. In *Network and Distributed Systems Security (NDSS) Symposium 2019*. https://doi.org/10.14722/ndss.2019.23351
[54] Twitter. 2019. https://ads.twitter.com/.
[55] Deepak Vasisht, Anubhav Jain, Chen-Yu Hsu, Zachary Kabelac, and Dina Katabi. 2018. Duet: Estimating User Position and Identity in Smart Homes Using Intermittent and Incomplete RF-Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 84 (July 2018), 21 pages. https://doi.org/10.1145/3214287
[56] Raghav H. Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-User Gesture Recognition Using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*. ACM, New York, NY, USA, 401–413. https://doi.org/10.1145/3210240.3210335
[57] Raghav H. Venkatnarayan and Muhammad Shahzad. 2018. Gesture Recognition Using Ambient Light. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 40 (March 2018), 28 pages. https://doi.org/10.1145/3191772
[58] Aditya Virmani and Muhammad Shahzad. 2017. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*. ACM, New York, NY, USA, 252–264. https://doi.org/10.1145/3081333.3081340
[59] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. 2018. Multi-Touch in the Air: Device-Free Finger Tracking and Gesture Recognition via COTS RFID. In *Proc. of IEEE INFOCOM*.
[60] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 851–860. https://doi.org/10.1145/2984511.2984565
[61] Wei Wang, Alex X. Liu, and Muhammad Shahzad. 2016. Gait Recognition Using Wifi Signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 363–373. https://doi.org/10.1145/2971648.2971670
[62] Xiao Wang, Tong Yu, Ming Zeng, and Patrick Tague. 2017. XRec: Behavior-Based User Recognition Across Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 111 (Sept. 2017), 26 pages. https://doi.org/10.1145/3130975
[63] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID Signal Reflection for Contact-free Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 193 (Dec. 2018), 22 pages. https://doi.org/10.1145/3287071
[64] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. Serendipity: Finger Gesture Recognition Using an Off-the-Shelf Smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3847–3851. https://doi.org/10.1145/2858036.2858466
[65] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E. Starner, Omer T. Inan, and Gregory D. Abowd. 2017. FingerSound: Recognizing Unistroke Thumb Gestures Using a Ring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 120 (Sept. 2017), 19 pages. https://doi.org/10.1145/3130985
[66] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-grained Hand Poses Using Active Acoustic On-body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 437, 10 pages. https://doi.org/10.1145/3173574.3174011
[67] Xiang Zhang, Lina Yao, Salil S. Kanhere, Yunhao Liu, Tao Gu, and Kaixuan Chen. 2018. MindID: Person Identification from Brain Waves Through Attention-based Recurrent Neural Network. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 149 (Sept. 2018), 23 pages. https://doi.org/10.1145/3264959
[68] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2018. PPG-based Finger-level Gesture Recognition Leveraging Wearables. In *Proc. of IEEE INFOCOM*.
[69] Zijie Zhu, Xuewei Wang, Aakaash Kapoor, Zhichao Zhang, Tingrui Pan, and Zhou Yu. 2018. EIS: A Wearable Device for Epidermal American Sign Language Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 202 (Dec. 2018), 22 pages. https://doi.org/10.1145/3287080

A DATASET DETAILS

The following URLs were used to retrieve each dataset:

- Walking Activity - https://archive.ics.uci.edu/ml/datasets/User+Identification+From+Walking+Activity
- NBA Player Statistics - https://www.kaggle.com/drgilermo/nba-players-stats
- CAT Scan Localization - https://www.kaggle.com/uciml/ct-slice-localization
- Activity Recognition - https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
- EEG Readings - https://www.kaggle.com/broach/button-tone-sz

A.1 Activity Recognition - A Common Mobile Platform Example

The initial purpose of this dataset was to determine the posture of a participant [6]. The authors recruited participants between the ages of 19 and 48 to provide an inertial measurement unit (IMU) data. The feature space consisted of time and frequency domain components read from a mobile phone based accelerometer and gyroscope. The full feature space has 561 features. This dataset is a typical example of the kinds of measurements that would be used to build a mobile platform identification system (e.g. gait recognition [19])

To process the dataset for use as an identification dataset, we merged the train and test sets that were provided with the ground truth labels, which we used as an additional feature. We then separated the user field from the measurements and used this as our class label. We then randomly selected 20 random labels and ran the multi-class classifiers.

The best confusion matrix for this dataset (see Figure 3) shows that all participants were easily distinguished by the random forest algorithm (black squares along the main diagonal of the matrix). This metric might lead us to believe that a system constructed from this data might be a viable identification system. Unfortunately, because the participant count is small, we do not have any idea how this system would perform when more participants were tested.

The feature list for this data is stored in a separate file names features.txt, there is a total of 561 features, all of which were used as the data. The user ID for each record was also stored in a separate file named subject_train.txt and subject_test.txt.

A.2 Walking Activity - Having a Small Number of Features Does Not Imply All Is Lost

The Walking Activity dataset was initially collected to perform gait authentication using two staged approaches. The dataset authors first inferred the posture from the data (e.g. walking or standing) and then perform a one class classification to determine if the posture readings correspond to the authorized user [10]. This dataset has only 4 features: time-step, x acceleration, y acceleration, z acceleration. There were a total of 22 participants who were all recruited by convenience methods.

To use this dataset to build an identification system, each of the individual participants readings was separated into individual files. To process this dataset we merged all of the files while applying a label corresponding to the file name/participant identifier. Since this data was time series, it would normally be the case that multiple samples would be analyzed as a group (e.g. within a time window of 5 seconds) to determine if they belonged to a particular participant. To see how far we could get with a naive approach, we treated each sample as a distinct measurement labeled by userID.

The random forest classifier had the best performance on this dataset, achieving $\approx 71\%$ accuracy (shown in Figure 3). While the identification performance of this dataset was not as good as the activity recognition dataset, it should be noted that this dataset is using significantly fewer features ($561 \rightarrow 4$). Since this is time series data, by treating each sample as distinct, we are not taking advantage of the temporal correlations that exist between samples. We could potentially capture some of these correlations if we folded samples that span a fixed interval into higher dimension samples. This would lower the total sample size but increase the feature space dimension,

which could capture some of the time information. Still, the naive approach does perform better than guessing, even though the temporal information is ignored. The results of this dataset demonstrate that even with a very low dimension feature space high performance can be achieved.

For the walking activity dataset all features were used. The id's served as labels.The feature list was: "time-step", "x acceleration", "y acceleration", "z acceleration".

A.3 CAT Scan Localization - When Discernibility Is High, More Testing Is Required to Identify the Limits of the System

This dataset consisted of 384 features extracted from full body CT scan images which were used to localize CT slices [23]. Data from 97 participants was analyzed and histograms of the physiological features (bone structures and air inclusions) were extracted. This dataset did not have an existing label, however, each record was labeled with a patientID. To process this dataset, we separated the patientID for each measurement and used it as the label. Because this dataset had such a large participant count we ran the analysis with several different 20 participants subsets. The support vector machine performance was 100% for most subset chosen (see Figure 3) and the variation between subsets was $\leq 1\%$.

The discernibility of this feature space is incredibly high. We ran the random forest algorithm with a 90 participant subset and only saw a $\approx 1\%$ drop in accuracy. To ensure that the features were not leaking label information into the classifier (e.g. one of the features may have been equal to the patientID), we identified the maximum feature importance (as reported by the random forest algorithm) and then removed all features that had an importance within 50% of that maximum (8 features total). With the most important features removed, the random forest classifier still achieved $\approx 99\%$ classification accuracy with a participant subset of 20, and $\approx 98\%$ accuracy with a subset of 90. The high discernibility of this feature space is not surprising since the measurements are the result of an entire body scan in a room sized instrument. Because the discernibility of this feature space is very good, determining an upper-bound on number of distinct individuals this type of measurement could distinguish would require testing with a significantly larger number of participants.

All features were used from this dataset. The patient_id served as the label, and all other columns were used as the feature vector. To ensure that one of the values was not highly correlated with the patient_id, we took the top 5 features from the high accuracy random forest and removed them from the dataset. When this was done the performance values did not change much.

A.4 EEG Readings - Poorly Discernible Features Will Easily Fail But PCA Might Give You Insight into Why

In this dataset measurements from a head mount EEG instrument were collected. There was a total of 81 participants, several of these who were suffering from Schizophrenia, a chronic illness. The original authors used the measurements to determine if there was a correlation between the illness and certain patterns in the measurements [17].

The EEG dataset represents measurement data where all users look very similar. In Figure 3 we show the best achieved classification results. The best algorithm was random forest however, the difference between random forest and neural networks was not significant (see Table 3). As in the CT scan case, we selected several random subsets of the participant count, and re-ran the analysis several times. Each sampling produced approximately the same results (no greater than $\approx 3\%$ variation).

Since the procedure and algorithms used for this dataset mirror the procedure and algorithms used for the CT scan dataset, why was the performance so different? One easy observation is that the size of the feature space is significantly smaller ($384 \rightarrow 40$). However, in the walking activity dataset, the feature space was significantly smaller than the activity recognition dataset, yet this system achieved $\approx 70\%$ accuracy.

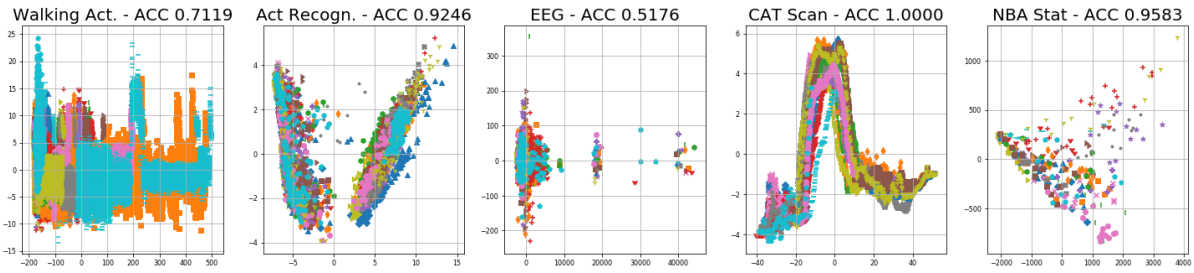

Fig. 9. The first two principal components of the measurements for each dataset is shown. These components can be used to explain the differences in performance observed for each dataset. For the CT scan, activity recognition and NBA stats datasets, the high accuracy is coupled with significant separation in the measurements from each participant. Distinct clusters can be observed indicating that even the limited information captured in these two components is enough to distinguish may participants. The walking activity dataset begins to hint at why performance would degrade as there are clusters, but they fail to be distinct and have some significant overlap. The EEG dataset demonstrates the worst case where there is almost no separation. Most of the participants measurements are overlapping in one large area with only a small portion lying outside this giant component. All test set measurements from each of the 20 participants that were classified in Figure 3 are plotted. The measurements form each participant have a distinct colors and marker shapes.

We can gain some insight into why the performance is so different by looking at a scatter plot of the first two principal components. In Figure 9, we show the first two principal components for each of the 20 participant subset that corresponds to the confusion matrix of Figure 3. As we can see, in the CT scan case, measurements from each participant form clusters and the overall shape are not uniformly distributed about the origin. In contrast, consider the EEG components where most of the measurements form one giant component sitting on top of the origin with other highly overlapping components to the right. This measurement dataset represents the worst case scenario for discernibility of the feature space. All participants produce measurements that vary significantly and the ranges over which they vary completely overlap. In this situation, it becomes very difficult to identify which measurement came from which participant.

For the EEG dataset we filtered out the following features:

u'trial', u'condition', u'ITI', u'rejected', u'Fz_N100', u'FCz_N100', u'Cz_N100', u'FC3_N100', u'FC4_N100', u'C3_N100', u'C4_N100', u'CP3_N100', u'CP4_N100', u'Fz_P200', u'FCz_P200', u'Cz_P200', u'FC3_P200', u'FC4_P200', u'C3_P200', u'C4_P200', u'CP3_P200', u'CP4_P200', u'Fz_B0', u'FCz_B0', u'Cz_B0', u'FC3_B0', u'FC4_B0', u'C3_B0', u'C4_B0', u'CP3_B0', u'CP4_B0', u'Fz_B1', u'FCz_B1', u'Cz_B1', u'FC3_B1', u'FC4_B1', u'C3_B1', u'C4_B1', u'CP3_B1', u'CP4_B1'

to eliminate some empty columns. The subject column was used as the label.

A.5 NBA Player Statistics - Large Numbers of Samples Per Participant Are Not Necessary If the Features Are Highly Discernible

The NBA Player statistics dataset is an artificial dataset constructed from player statistics spanning the years 1950 to 2017 [21]. Since the player careers tend to be shorter than 25 years, this dataset was explicitly constructed to have only a small number of samples per players on average. The stats of a player is a function of the players' capability and environmental circumstances of that year. Thus each measurement's features should be centered around a mean, but have random variation due to environmental factors (e.g. health or number of home games).

The dataset was processed to have a small number of partially distinct samples from each player, but a large number of players overall. The player statistics served as a measurement source with enough variation such that

Table 5. All parameter arguments used for every algorithm, dataset pair

	Act. Recogn.	Walking Act.	CT Scan	EEG	NBA Stat.
RandomForestClassifier	`n_estimators =1000 max_depth =20`	`n_estimators =1000 max_depth =20`	`n_estimators =1000 max_depth =20`	`n_estimators =1000 max_depth =20`	`n_estimators =1000 max_depth =20`
svm.SVC	default	default	default	default	default
MLPClassifier	default	default	default	default	default

there should be some overlap between players. Each measurement of a player has the potential to overlap with another player's measurement. When this happens the classifier may confuse one player for another.

All players were given a unique numerical identifier, and each time-stamped playing statistic was treated as a single measurement. Several categorical values (E.g. team or position) were also encoded as a single integer value. Only players with at least 15 measurements were counted, thus the final dataset had measurements from 290+ players (there are very few datasets with such a large participant size).

We conducted the same analysis on this dataset as on all other previous datasets. Given the small measurement size and low density of measurements per participant, as compare to all other datasets, we might expect much worse performance. The classification performance, however, is high because the each player has values that are very distinct. The team and position alone provide significant clustering into distinct groups. These groups are then refined further by playing characteristics. In Figure 9 we see the first two principal components of all the samples in the test set for the 20 players shown in the CM of Figure 3. Even though the point density is very low ($\approx$ 15 per player), the clustering of each players samples is very tight. These two features alone provide significant discernibility between players.

For the NBA player statics, all names were encoded as numeric identifiers including the player name and team name. The full feature list was:

u'Year', u'Player', u'Pos', u'Age', u'Tm', u'G', u'GS',u'MP', u'PER', u'TS%', u'3PAr', u'FTr', u'ORB%', u'DRB%', u'TRB%',u'AST%', u'STL%', u'BLK%', u'TOV%', u'USG%', u'blanl', u'OWS', u'DWS',u'WS', u'WS/48', u'blank2', u'OBPM', u'DBPM', u'BPM', u'VORP', u'FG',u'FGA', u'FG%', u'3P', u'3PA', u'3P%', u'2P', u'2PA', u'2P%', u'eFG%',u'FT', u'FTA', u'FT%', u'ORB', u'DRB', u'TRB', u'AST', u'STL', u'BLK',u'TOV', u'PF', u'PTS'

Where player was the player name which served as the label to be predicted.

B PARAMETERS USED FOR EACH ALGORITHM ON EACH DATASET

Table 5 enumerates all the parameters used for each dataset. In most cases the defaults were sufficient. For the case of EEG we used a random parameter search to try to improve results for the Random Forest and Support Vector Machine algorithms, but there was no significant gain in performance. The default neural network has a single hidden layer with 100 neurons. The default kernel for support vector machines was the RBF kernel and it makes multi-class decisions via one-vs-one run offs.